Company X is a tech-led company that connects informal retailers and other similar outlets directly to fast-moving consumer goods companies (FMCGs) by communicating demand from retail outlets directly to FMCGs and their distributors and managing the delivery of the goods to the retailers.

Question 1: Dynamic Routing

Dynamic routing is an order fulfilment strategy where orders are dynamically grouped to create an optimum delivery route. An optimum delivery route is one that will cost the least and ensures that the vehicle is optimally utilized.

Assuming delivery costs are broken down as below;

⎯ Fuel costs = Distance (Kms) covered \* Fuel cost per KM

⎯ Vehicle costs = cost per trip \* no of trips

To create optimum routes, it therefore requires one to ensure that the vehicle covers the shortest distance and carries the most weight per trip.

Remember, per trip, a vehicle will carry a maximum of its capacity (vehicles have a standard carrying capacity as per the manufacturing specifications).

To achieve a dynamic route;

Per territory, per delivery window, select a combination of orders that will cover the shortest distance but provides the highest utilization. (Utilization = Vehicle capacity (as per manufacturer)/total weight of scheduled orders. There is also a consideration of the capacity in terms of space occupied by a product.

In this case study;

You are only required to address one element of dynamic routing which is clustering shops (customer\_id) around the fulfilment centers.

You will submit an output and the algorithm you will use to cluster the shops. Plot the output on the map and share screenshots.

Check the sample

⎯ Q1\_sample\_submission\_file\_centroids for the columns required and

⎯ Q1\_sample\_submission\_file\_screenshots for screen shots of the clusters and the algorithm used.

Question 2: Churn Prediction

Please prepare a presentation (and csv for the second task) that includes the answers to the tasks described below. You’ll find more information for the tasks attached.:

1. Using the attached dataset (dataset.csv) analyze the data and visualize the most important aspects using your preferred method. Furthermore, share three ideas on how to decrease the churn rate. Document your steps where needed.

2. Split the data into train and test sets. Predict whether a shop will churn or not. Please document your steps and method used. The csv, “Q2\_sample\_submission\_file\_churn” will help with the format.

The objective of this case study is to create a machine learning model that will predict whether a customer will churn or not.

Question 3: Smart pricing

Given that we use digital platform to receive orders from customers and deliver/fulfil, is impossible to conduct real RCT or A/B tests providing customers we know are similar with completely different prices. Given this, and given that Kenya is divided into 19 Territories, how would you structure an experiment that follows the principles of a RCT and how would you assess the level of confidence and the margins of error associated with it?

Please prepare a 3 slides presentation on your approach

Given the constraints of not being able to conduct real RCTs or A/B tests with different prices for similar customers due to the nature of your digital platform and the regional segmentation in Kenya, you can still design a structured experiment to evaluate pricing strategies. Here's a detailed approach to structuring such an experiment and assessing the level of confidence and margins of error:

### 1. \*\*Segment the Experiment by Territories\*\*

Since Kenya is divided into 19 Territories, you can use these territories as the primary units of randomization. The idea is to apply different pricing strategies in different territories while keeping the customer characteristics similar within each territory. This approach leverages the geographic segmentation to approximate the principles of randomization in an RCT.

#### Steps to Structure the Experiment:

\*\*a. Define Pricing Strategies:\*\*

- \*\*Control Group:\*\* Standard pricing.

- \*\*Treatment Groups:\*\* Various alternative pricing strategies (e.g., discount pricing, premium pricing, bundling).

\*\*b. Randomize Pricing Strategies Across Territories:\*\*

- Randomly assign different pricing strategies to different territories. Ensure that each pricing strategy is applied to a sufficiently large number of territories to ensure statistical power.

\*\*c. Ensure Similarity Within Territories:\*\*

- While randomizing pricing strategies, make sure the territories are comparable in terms of key customer characteristics and market conditions. This might include factors such as average income, customer demographics, or order volume.

\*\*d. Implement Pricing Strategies:\*\*

- Roll out the pricing strategies in the designated territories, ensuring that each territory's pricing remains consistent throughout the experiment.

### 2. \*\*Data Collection and Metrics\*\*

\*\*a. Collect Data:\*\*

- Gather data on sales performance, customer behavior, and financial metrics from each territory. Key metrics might include average order value, total revenue, customer acquisition cost, and customer retention rates.

\*\*b. Measure Effectiveness:\*\*

- Track performance against your objectives (e.g., revenue growth, profitability) and compare these metrics across different pricing strategies.

### 3. \*\*Analyze the Results\*\*

\*\*a. Compare Groups:\*\*

- Use statistical methods to compare the outcomes of different pricing strategies. Common statistical tests include ANOVA (Analysis of Variance) or regression analysis to determine if there are significant differences between the control and treatment groups.

\*\*b. Account for Confounders:\*\*

- Control for any potential confounding variables that might affect the outcome (e.g., regional economic differences, seasonality). This can be done using multivariate regression techniques.

\*\*c. Assess Statistical Significance:\*\*

- Evaluate the statistical significance of your results to determine if observed differences are unlikely to be due to random chance.

### 4. \*\*Assess Confidence and Margins of Error\*\*

\*\*a. Determine Sample Size:\*\*

- Calculate the required sample size for each territory based on the expected effect size, desired power (e.g., 80% or 90%), and significance level (e.g., 0.05). This ensures that your results are statistically reliable.

\*\*b. Estimate Confidence Intervals:\*\*

- Calculate confidence intervals for your key metrics to understand the range within which the true effect is likely to lie. Wider intervals indicate more uncertainty.

\*\*c. Conduct Power Analysis:\*\*

- Perform a power analysis to determine the probability of detecting a true effect if it exists. This helps in understanding the robustness of your experiment.

\*\*d. Check for Internal Validity:\*\*

- Verify that the experiment was conducted consistently across territories and that no external factors introduced bias.

### Example:

Assume you have assigned four different pricing strategies (including the control) to four different groups of territories. After a set period, you analyze the sales data and find that the average revenue per order significantly differs between the pricing strategies. You would then calculate the statistical significance of these differences and assess the confidence intervals to determine the precision of your estimates.

### Conclusion:

By using territorial randomization and ensuring that territories are similar in key aspects, you can approximate the conditions of an RCT. Statistical analysis will help you assess the effectiveness of different pricing strategies and provide confidence in your results. This approach allows you to make informed decisions about pricing while accounting for the limitations of your digital platform and regional constraints.

About the Data

Dataset 1 (dataset.csv)

The dataset provided includes order & delivery details. The source is the Kyosk Digital platform.

Files available for download:

⎯ dataset.csv – contains customer orders and delivery data. This data set will be used for both Q1 and Q2. To answer Q2, split the data to train and test sets.

The order value is the net value per order\_id

Column

Description

order\_id: uuid of an order

customer\_number: unique customer/shop id

transaction\_date: date of order

description: Items ordered by the customer

total\_qty: no of units ordered

net\_total: net total for the order

loyalty\_amount: amount for loyalty points

loyalty\_points: no of loyalty points

discount\_amount: amount for the discount

customer\_group: customer type

territory: territory where the customer is registered/served from

delivery\_date: date of delivery

delivery\_window\_end\_time: time the order is to be delivered (start)

delivery\_window\_start\_time: time the order is to be delivered(end)

set\_warehouse: warehouse to deliver from

duka\_latitude: latitude of the duka

duka\_longitude: longitude of the duka

⎯ territories\_centroids.csv – contains the geo location of all territory warehouses/fulfillment centers which will be the central place in your clusters (centroids).

Use this dataset to answer Q1.

Column- Description

Territory- territory where the customer is registered/served from

warehouse\_latitude-latitude of the warehouse (Cluster centroid)

warehouse\_longitude-longitude of the warehouse (Cluster centroid)